Project Testbed: Argument Mapping and Deliberation Analytics

Other

Project Acronym:	**CATALYST**
Project Full Title:	**Collective Applied Intelligence and Analytics for Social Innovation**
Grant Agreement:	**6611188**
Project Duration:	**24 months (Oct. 2013 - Sept. 2015)**

D4.2b Project Testbed: Argument Mapping & Deliberation Analytics

Deliverable Status:	**Final**
File Name:	**CATALYST_ D4.2b.pdf**
Due Date:	**September 2015 (M24)**
Submission Date:	**November 2015 (M26)**
Dissemination Level:	**Public**
Task Leader:	**Purpose**

This project has received funding from the European Union's Seventh Framework Programme for research, technological development and demonstration under grant agreement n°6611188

The CATALYST project consortium is composed of:

SO	Sigma Orionis	France
I4P	Imagination for People	France
OU	The Open University	United Kingdom
UZH	University of Zurich	Switzerland
CSCP	Collaborating Centre on Sustainable Consumption and Production	Germany
Purpose	Purpose Europe	United Kingdom
Wikitalia	Wikitalia	Italy

The CATALYST project has received funding from the European Union's Seventh Framework Programme for research, technological development and demonstration under grant agreement n°6611188

Revision Control

Version	Author	Date	Status
0.1	Marc-Antoine Parent	October 30, 2015	Initial Draft
0.2	Anna De Liddo, Thomas Ullmann	November 6, 2015	Inputs from the Open University
0.3	Mark Klein	November 10, 2015	Inputs from the University of Zürich
0.4	Marc-Antoine Parent	November 19, 2015	Draft
0.5	Marta Arniani	November 23, 2015	Quality Check
0.6	Marc-Antoine Parent	November 26, 2015	Final Draft reviewed
1.0	Marta Arniani	November 26, 2015	Submission to the EC

The CATALYST project has received funding from the European Union's Seventh Framework Programme for research, technological development and demonstration under grant agreement n°6611188

Table of Contents

Executive summary 5

Introduction 6

1. Analytics and alerts implemented 7

1.1 Metrics 7
1.2 Alerts 12
1.3 Analytics Identification Methodology 15
1.4 Visual analytics based on attention metrics 18
1.4.1 Activity bias visualisation 18
1.4.2 Rating bias visualisation 19
1.4.3 Attention map visualisation 19
1.4.4 Community interest network visualisation 20
1.4.5 Sub-communities network visualisation 21
1.5 Alerts and user feedback interface in LiteMap and DebateHub 22
1.5.1 Alerts in LiteMap 23
1.5.1.1 LiteMap's alert interface 23
1.5.1.2 User feedback Interface 24
1.5.2 Alerts in DebateHub 24
1.6 Semantic clusters 26

2. Community testing of visual analytics 26

2.1 Evaluation 27
2.1.1 Background information 27
2.1.2 Task performance 27
2.1.3 Usability 27
2.1.4 Discussion 28

3. Community testing of alerts 28

3.1 Loomio test of harvester alerts 28
3.1.1 Experimental design 29
3.1.2 Testbed outcomes 29
3.2 Seventh Sustainable Summer School 30
3.2.1 Participation-centric analytics 30
3.2.2 Semantic clusters and semantic proximity 31

4. Post-hoc testing 31

4.1 Alerts 31
4.2 Metrics and visualization 32
4.3 Suggestions based on semantic clusters 37

Conclusions and future work 40

References 41

List of Figures 42

List of Tables 42

The CATALYST project has received funding from the European Union's Seventh Framework Programme for research, technological development and demonstration under grant agreement n°6611188

Executive summary

The present document is a deliverable of the CATALYST project, funded by the European Commission's Directorate-General for Communications Networks, Content & Technology (DG CONNECT), under its 7th EU Framework Programme for Research and Technological Development (FP7).

One key goal of the Catalyst project was to design metrics that could capture and represent aspects of the conversation's structural quality, to assist harvesters and moderators. Many such metrics, alerts and visualizations were developed in the course of the project, but initial user testing has shown that users find it difficult to interpret abstract signals. Following that, we have both introduced new analytics that we felt could be more directly useful, and improved the representation of existing ones. We attempted to test those later refinements, but could not do it with large communities, as planned. Instead, we evaluated their usefulness in a smaller conversation and in experimental settings.

The CATALYST project has received funding from the European Union's Seventh Framework Programme for research, technological development and demonstration under grant agreement n°6611188

Introduction

Discussion analytics aim to represent patterns in the conversation processes that might be indicative of various community functions or dysfunctions. The CATALYST analytics server is a free open-source **web service** that deliberation platforms can use to address these important challenges. It provides a substantive library of analytics for assessing crowd-authored deliberation maps. These analytics take two forms: metrics and alerts:

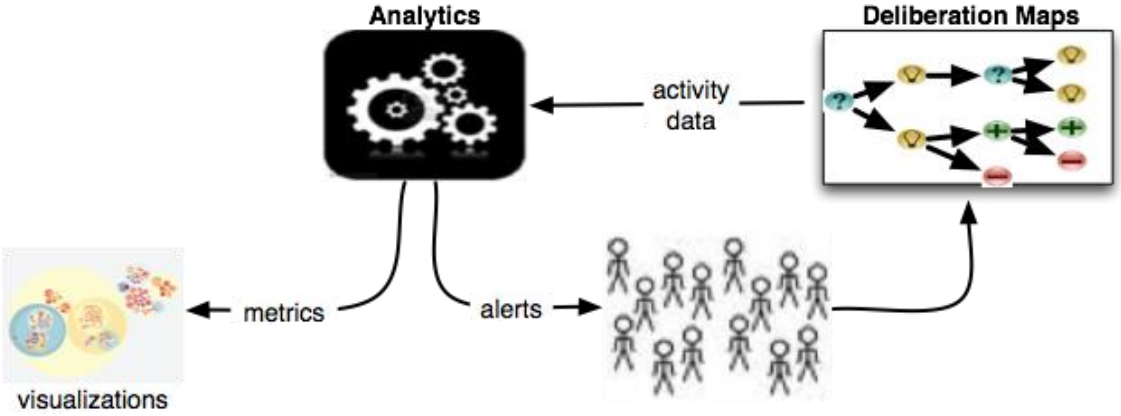


Figure 1 - Metric and alert server data flow

The role of *metrics* is to provide summary overviews of the status of the deliberation, highlighting phenomena (such as controversy hot spots or balkanization) that would be difficult to identify simply by browsing the map on a post-by-post basis. Visualization based on the metrics can enable community managers (moderators and harvesters) to monitor the health of the conversation, and by extension, they can also make participants in general aware of conversation patterns. Some of those metrics can be developed into *alerts*, which will signal to participants that some of the metrics' scores have passed certain thresholds that require attention, and also provide user-specific notifications of what elements of the deliberation map a participant should consider in order to maximize their contribution.
In particular, analytics and alerts allow the following:

- Help community managers monitor patterns of participation (either decaying or balkanized participation);
- Help participants find topics and other participants that could interest them;
- Help harvesters find emerging topics;
- Help moderators find various communication dysfunctions.

The analytics implemented in this server were *identified* based on systematic analysis of deliberation needs and challenges, *prioritized* by CATALYST team members using a crowd-sourced online tool, *implemented* using state-of-the-art data mining techniques such as social network analysis, dimensionality reduction, and linguistic corpus analysis, and *evaluated* using a range of empirical tests.

We have integrated those tools into our deliberation platforms, and they were put in the hands of community managers and participants. From there, we gathered their responses to the visualizations and alerts at their disposal.

The CATALYST project has received funding from the European Union's Seventh Framework Programme for research, technological development and demonstration under grant agreement n°6611188

1. Analytics and alerts implemented

1.1 Metrics

The role of metrics is to provide easy-to-understand summaries of different aspects of the health of a deliberation engagement. While in some cases the metrics are single numbers that can be presented as is, in many cases the metrics return a table of numbers that need to be rendered using a simple visualization, such as a scatter plot, in order to make their message clear. The visualizations developed by the CATALYST project are described in a separate part of the CATALYST final report, and recapitulated below.

The following table summarizes the metrics implemented in the CATALYST analytics server. The table gives the name of each metric and describes how the metric works:

Table 1. List of metrics

Name	Description
branchsize	Returns the product of the average width and depth for the posts in a branch.
controversy	Returns the controversy score for every post in a given branch, based on the ratings the posts got.
user activity	Returns data about the user activity to date for a given branch.
interest by user	Measures the level of interest for each user in each post.
interest by community	Measures the level of *community* interest for each post in the map.
interest inequality	Measures to what extent community interest been applied *unequally* to the children of each post, where 0 = fully equal, 1 = all discussion is on single child. No data is given for nodes with zero or one children, since the inequality is zero, by definition, in those cases.
interest space dimensions	Defines the topics in the discussion, where a topic is a set of posts that all tend to be looked at together. The tweight (a number) specifies how useful the topic is for clustering user activity. The pweights specify how important the corresponding posts are in each topic. To visualize this, you can display a version of the argument map which shows only the high-weighted posts for a given topic, with a font size that is proportional to the pweight.
interest space post coordinates	Gives how active each post is in each topic. People tend to be interested in one post tend to also be interested in other posts with similar coordinates.
interest space post clusters	Returns clusters of posts that tend to be looked at together. This metric uses singular vector decomposition to find posts whose activity is correlated. So, for example, we may find people who pay a lot of attention to the posts on stock markets also typically pay a lot of attention to the posts on banking. Posts with highly correlated activity can then be said to represent a “theme” or “topic” family. In the example above, the theme would be something like “finance”. So, this metric looks for topics, and tells you all the posts in that topic.

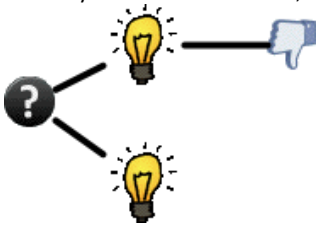
The CATALYST project has received funding from the European Union’s Seventh Framework Programme for research, technological development and demonstration under grant agreement n°6611188

	This can be used to give recommendations (e.g. look at the other posts in a cluster if you looked at one of the posts) as well as to reveal dependencies across issues in a map (i.e. if posts from different issue branches have correlated activity).
interest space post clustering	This measures *how much* clustering occurs in the post, on a scale from 0 (no clustering) to 1 (clear distinct clusters).
interest space user coordinates	Gives the degree of interest each user has in each topic. Users with similar coordinates tend to have similar interests.
interest space user clusters	Identifies clusters of users who are interested in the same topics.This metric uses singular vector decomposition. It looks for correlations in what users attend to e.g. when people who look at the posts on the stock market almost always look at the posts on banking. The posts whose activity is highly correlated can then be viewed as representing a "theme" e.g. a "financial" theme, in the example above. Let's say our map ends up having two major themes: posts about finances, and posts about sports. And we classify people by their activity on each theme. We may find clusters e.g. a group of people who are interested in sports but not finance, and another group that is the opposite. Each cluster is a "community". So, the short answer is that a community is a group of users that have similar interests.
interest space user clustering	Returns a value, from 0 to 1, that measures *how much* clustering occurs in the users.A high degree of clustering means that there are distinct groups (i.e. balkanization).
interest narrowing	Specifies how *quickly* (over time) the attention focused on a subset of the children of a post.
support by user	Returns the rating the users gave each post, on a scale of 1 (low) to 5 (high).
support by community	Measures the level of *community* support for the posts in a map, factoring in support for the posts below it, on a scale of 1 (low) to 5 (high). ● an issue's importance is the average of it's own ratings and those of its sub-issues ● an idea's promise is the average of it's own ratings and those of its sub-ideas and arguments ● an argument's convincing-ness is the average of it's own ratings and those of its arguments
support inequality	Measures to what extent the community support is unequal forthe ideas under an issue, where 0 = fully equal, 1 = all discussion is on single child. NB: data is only given for issues with ideas attached. Calculated using a gini coefficient.
support space dimensions	Describes the biases in the discussion, where a bias = a set of posts that tend to be liked together. The bweight (a number) specifies how useful the bias is for clustering user ratings. The pweights specify the importance of the most important posts in each bias. If the branch root is not given, this metric looks at the entire discussion. To visualize this, you can display a version of the argument map which shows only the high-weighted posts for a given topic, with a font size that is proportional to the weight.

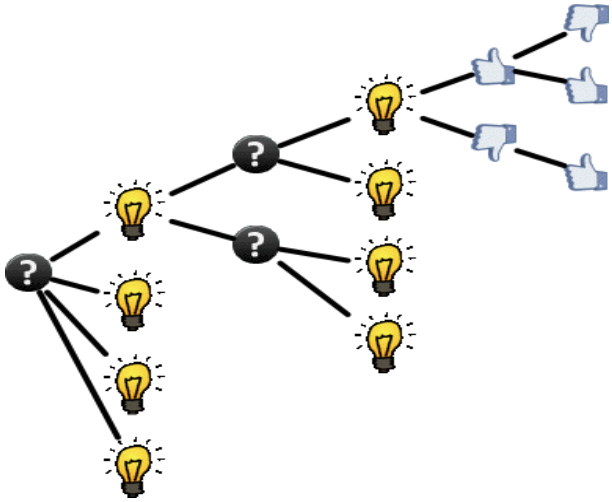
The CATALYST project has received funding from the European Union's Seventh Framework Programme for research, technological development and demonstration under grant agreement n°6611188

support space post coordinates	Gives the support space coordinates for all the ideas in the given branch. A bias is a set of posts whose ratings tend to be correlated. For example, if people who like solar also tend to like wind power, this would represent a "bias" for both forms of renewable energy. People who like one post tend to also like other posts with similar support space coordinates. If the branch root is not given, this metric looks at the entire discussion.
support space post clusters	Returns clusters of posts that tend to be liked together.
support space post clustering	Returns a value, from 0 to 1, that measures *how much* clustering occurs in the support space for posts.
support space user coordinates	Gives a user's position in the support space. Users with similar coordinates have similar biases. A highly balkanized or polarized discussion will result in distinct clusters of users in the support space.
support space user clusters	Returns clusters of users with similar biases.
support space user clustering	Returns a value, from 0 to 1, that measures *how much* clustering occurs in users based on their biases.
map topology	Gives topology statistics for posts in the map. If "root" is not specified, it looks at the entire map. The fields include: type-issue | idea | pro | con numposts-number of posts in the branch under the post numissues-number of issues in the branch numideas-number of ideas in the branch numpros-number of pro arguments in the branch numcons-number of con arguments in the branch mindepth-minimum depth of posts in the branch maxdepth-maximum depth of posts in the branch avgdepth-average depth of posts in the branch stdevdepth-standard deviation of posts in the branch minbreadth-minimum breadth (# of children) in the branch (excludes leaves, whose breadth is 0) maxbreadth-maximum breadth (# of children) in the branch (excludes leaves, whose breadth is 0) avgbreadth-average breadth (# of children) in the branch (excludes leaves, whose breadth is 0) stdevbreadth-standard deviation of breadth (# of children) in the branch (excludes leaves, whose breadth is 0) **only for issue or idea posts** minsdepth-minimum depth of "skeleton" (issue and idea) posts in the branch maxsdepth-maximum depth of "skeleton" (issue and idea) posts in the branch avgsdepth-average depth of "skeleton" (issue and idea) posts in the branch stdevsdepth-standard deviation of depth of "skeleton" (issue and idea) posts in the branch

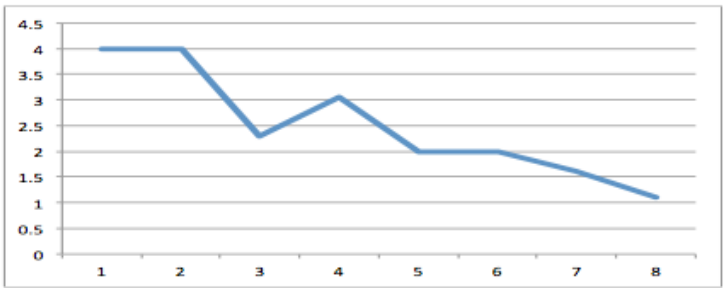
The CATALYST project has received funding from the European Union's Seventh Framework Programme for research, technological development and demonstration under grant agreement n°6611188

	minsbreadth- minimum breadth of skeleton posts in the branch (excludes leaves, whose breadth is 0) maxsbreadth- maximum breadth of skeleton posts in the branch (excludes leaves, whose breadth is 0) avgsbreadth- average breadth of skeleton posts in the branch (excludes leaves, whose breadth is 0) stdevsbreadth- standard deviation of breadth of skeleton posts in the branch (excludes leaves, whose breadth is 0) -**only for idea posts** minadepth- minimum depth of argument (pro con) tree under the idea maxadepth- maximum depth of argument (pro con) tree under the idea avgadepth- average depth of argument (pro con) tree under the idea stdevadepth- standard deviation of depth of argument (pro con) tree under the idea minabreadth- minimum breadth of argument (pro con) tree under the idea (excludes leaves, whose breadth is 0) maxabreadth- maximum breadth of argument (pro con) tree under the idea (excludes leaves, whose breadth is 0) avgabreadth- average breadth of argument (pro con) tree under the idea (excludes leaves, whose breadth is 0) stdevabreadth-standard deviation of breadth of argument (pro con) tree under the idea (excludes leaves) nsupports- number of arguments that support the idea (e.g. pros, cons of cons …) nauthorsupports-number of supporting arguments created by the author of the idea preachiness-nauthorsupports/nsupports: the fraction of supporting arguments that come from the idea author
expertise	Specifies the average rating for the posts a user has contributed in a given topic. 0 means the user didn't contribute anything to that topic.
controversy	Specifies how controversial the discussion for an issue is, ranging from 0 (low controversy) to 1 (highly controversial).
maturity	Specifies how mature the discussion for an issue is. This can be estimated easily, in deliberation maps, by gathering statistics on the *topology* of the branch (e.g. in terms of breadth and depth of issues, ideas and arguments) for that topic. The following issue, for example, probably represents a relatively immature discussion, because it includes few ideas and arguments:  The following issue, by contrast, represents a more mature discussion:

The CATALYST project has received funding from the European Union's Seventh Framework Programme for research, technological development and demonstration under grant agreement n°6611188

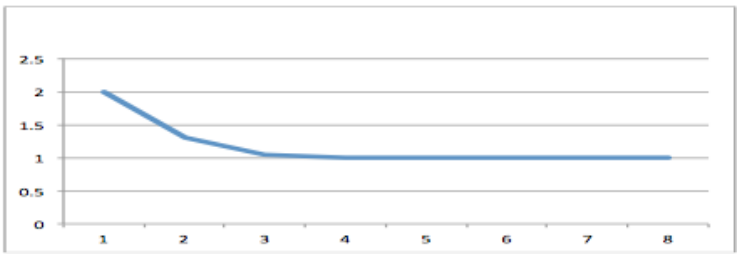
	
agreement	This returns a table, for each issue, where the rows and columns represent users, and the cells represent how much each user pair agrees about the best ideas for the issue. This could be visualized as a force-directed graph where nodes = users, where agree and disagree links have different colors, and where users that agree with each other are placed close to each other and far from those they disagree with. Balkanization and polarization would show up as strong clustering in the graph.
support consistency	Measures to what extent an idea's average rating is consistent with the ratings for the underlying arguments. This can be done for the ratings from an individual user, for a group of users, or for all users.
social graph	Returns a graph showing which users have interacted (i.e. have rated, commented on, responded to, or edited posts created by the other user). "linktype" currently just gives the number of links between the two users.
groupthink	Returns an estimate of the level of groupthink in the deliberation for a given issue. groupthink occurs when a crowd converges prematurely on a given (often the first) solution idea, without giving adequate attention to competing ideas. This can be detected at how quickly the Gini coefficient (which measures inequality) increases for the ideas addressing an issue in the deliberation map. Consider the following examples. In the first example, the participant's activity related to each competing idea (rendered at each time period as the size of the idea icon):

The CATALYST project has received funding from the European Union's Seventh Framework Programme for research, technological development and demonstration under grant agreement n°6611188



In this case, the gini coefficient declines slowly over time, indicating that the community considers a range of options for a while.

In the second example, one idea almost immediately becomes the sole focus of the community's interest, resulting in a gini coefficient that drops very rapidly.



1.2 Alerts

The role of alerts is to inform participants of opportunities that maximize their ability to contribute positively to a deliberation. An alert can point users to map posts they should view, as well as to other users they should be aware of. Alerts do so using information about the deliberation map as well as user roles and activity, thereby building a model of user interests and expertise

The CATALYST project has received funding from the European Union's Seventh Framework Programme for research, technological development and demonstration under grant agreement n°6611188

as well as of the deliberation maps' strengths and gaps. A matchmaking procedure then points users to the parts of the deliberation where they can do the most good.

Each alert is a kind of "daemon" that continually scans the deliberation map for instances that trigger it. Every alert has a "strength" value, representing how strongly it was triggered. Alert triggers are specified using a pattern query language (Klein and Bernstein, 2004) developed by a CATALYST team member. In addition to looking for such "local" data patterns as "user X rated post Y without viewing underlying argument Z", alerts can also look for extreme values in the summary metrics described in the previous section. Posts in a deliberation map may trigger multiple alerts. The deliberation system can therefore highlight, for each user, the posts that have multiple highly-weighted alerts associated with them e.g. using the following kind of user interface:

Figure 2 - UI for alerts in tree view

The following table summarizes the alerts implemented in the current version of the CATALYST analytics server. The table gives the name of each alert, indicates the information flow (i.e. what kind of entity is proposed to what kind of user[1]), and describes how the alert works. The currently implemented alerts include:

Table 2. List of alerts

Name	***Flow***	***Description***
unseen by me	post -> author	The author has not yet viewed the post
response to me	post -> author	The post is a response by someone else to a post created by the author.
unrated by me	post -> author	The author has not yet rated this post.
lurking user	user ->moderator	The user has not edited or created any posts

[1] We distinguish three types of user role: author (responsible for contributing and rating the issues ideas and arguments that make up a deliberation map), moderator (responsible for ensuring a deliberation engagement achieves useful results), and customer (responsible for defining the aims of the deliberation as well as for using its results).

The CATALYST project has received funding from the European Union's Seventh Framework Programme for research, technological development and demonstration under grant agreement n°6611188

ignored post	post -> author	The post has not been viewed by anyone but original author
mature issue	post -> customer	The issue has>= 3 ideas with at least one argument each. The "strength" of the alert is a function of the total size of the deliberation map branch under that issue. This alert can help *customers* find the parts of the deliberation map that are mature and thus ready to support a decision process.
immature issue	post -> moderator, author	The inverse of “mature issue”; i.e. the post is an issue which does NOT have at least three ideas with at least one argument each.
well evaluated idea	post -> moderator, customer	An idea post has several pros and cons, including some rebuttals
poorly evaluated idea	post -> author	The inverse of “well evaluated idea”; an idea post has few pros and cons, and no rebuttals
inactive user	post -> moderator	The user has been inactive in the deliberation.
interesting to me	post -> author	The post should interest a user, because it is close, in the deliberation map, to posts the author attended to in the past.
interesting to people like me	post -> author, customer	This post was viewed by people whose interests are similar to the user. This alert uses SVD (see section 5) to produce a low-dimensional rendition of what posts users attend to, and then uses cosine similarity to identify posts that are close in that space to posts that the user has already seen. The weight for this trigger is an inverse function of the cosine distance between a proposed post and posts that the user has already seen.
supported by people like me	post -> author, customer	This post was highly rated by people whose opinions are similar to the user. This operates the same as the "interesting to people like me" metric, only it is applied to rating scores rather than activity scores.
hot post	post -> author	This post is in the top quartile of the most active posts for the last 24 hours.
orphaned idea	post -> author	This idea post is receiving substantially less activity than it’s siblings.
winning idea	post -> author	This idea is receiving the predominance of community support for a given issue. NB: if all the ideas have a high support score, there is no clear winner so this alert will not fire.
contentious issue	post -> author	This is an issue with ideas that the community ratings are strongly divided over.
controversial idea	post -> author	The community has widely divergent opinions (as reflected by their ratings) of whether an idea is a good one or not.

The CATALYST project has received funding from the European Union’s Seventh Framework Programme for research, technological development and demonstration under grant agreement n°6611188

inconsistent support	post -> author	This is an idea where the user's support for the idea and for it's underlying arguments are inconsistent: propagating the support values from the arguments to the idea produce an inferred rating that is quite different than the rating the user gave the post. This alert can be symptomatic either of an ill-considered rating, or of the user evaluating an idea based on arguments that he/she has not rated or added to the map yet.
person with interests like mine	user -> author, customer	Identifies a user who has similar interests to me, based on activity patterns. The strength score reflects closeness of match.
person who agrees with me	user -> author, customer	Identifies a user who has similar opinions to mine, based on rating patterns. The strength score reflects closeness of match.
user gone inactive	user -> moderator	Identifies a user who was initially active, but have been inactive for at least a duration specified in the analytics request.
rating ignored argument	post -> author	Identifies a relevant argument that the user did not view before rating a post. Clearly, a careful rating for a post should take into account the arguments for and against it.
rating ignored competitor	post -> author	Identifies a competing idea (i.e. that responds to the same issue) that the user did not view before rating a post. This can be valuable because the scale on which we rate ideas will often be influenced by the overall strength of the competing ideas.
unseen response	post -> author	Identifies a response authored by someone else to a post I authored.
unseen competitor	post -> author	Identifies an idea authored by someone else that competes (by virtue of responding to the same issue) with an idea I authored.
user ignored competitors	user -> author, moderator	Identifies a user who ignored competitors to their ideas.
user ignored arguments	user -> author, moderator	Identifies a user who ignored underlying arguments when rating posts.
user ignored responses	user -> author, moderator	Identifies a user who ignored responses authored by other people to their posts.

In addition, the alert definition language makes it easy to define alerts that look for extreme values in summary metrics e.g. an issue where the balkanization metric value is high.

1.3 Analytics Identification Methodology

The analytics we implemented in the CATALYST server were identified using a systematic methodology, known as *process-goal-exception* analysis, that was developed by a member of the CATALYST team (Klein, 2012). The key idea is that analytics can be viewed as the tools we use to identify how well a process is achieving (or failing to achieve) its goals. In this method, we thus first define a normative model of the target process and what goals would ideally be achieved for each process step. For each goal, we

The CATALYST project has received funding from the European Union's Seventh Framework Programme for research, technological development and demonstration under grant agreement n°6611188

then identify analytics for measuring its success, as well as how that goal can be violated (the exceptions). For each exception, finally, we identify analytics for assessing when an exception is taking place. Each metric can be annotated with the kind of user that would be interested in that kind of information.

Identify normative process model: The first step is to identify a model of how the target process should work. The core process supported by the CATALYST ecology is "social innovation: i.e. crowd-based solution identification. Our model of this process includes the following subtasks:

Social Innovation Process			
Identify problems to solve	Identify possible solutions for these problems	Evaluate the candidate solutions	Select the best solution(s) from amongst the candidates

Identify goals: The next step is to identify what each task in the process should ideally achieve: its' goals. Our current model of the social innovation process includes the following goals:

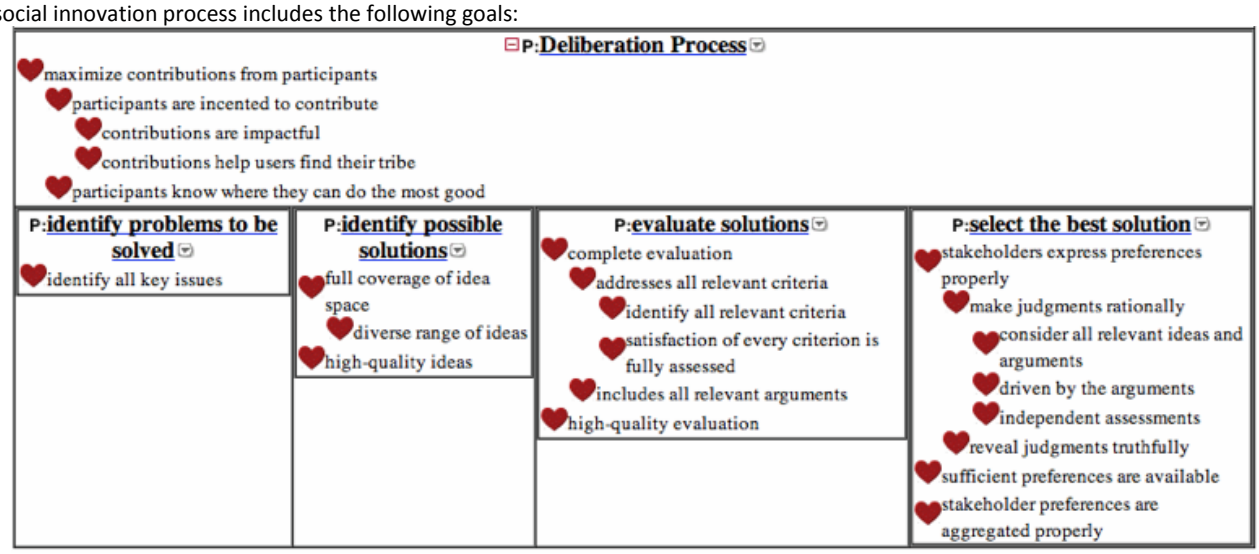


Figure 3 - Selection of goals for alerts

P: *= process,= goal.*

A social innovation process should, for example, use a good process (i.e. where the right people contribute actively and effectively to performing the most critical tasks) to achieve good results (i.e. complete, high-quality, well-organized content) while also strengthening and learning about the members of the user community.

Identify exceptions: For each goal, we then identify how it can be violated (the *exceptions*). The goal of having the right participants involved, for example, can have the following exceptions:

The CATALYST project has received funding from the European Union's Seventh Framework Programme for research, technological development and demonstration under grant agreement n°6611188

Figure 4 - Selection of exceptions for alerts

= *exception= metric process= handler process*

We can have too few authors, for example, or inadequate diversity in the author population.

Identify Analytics: For each exception, finally, we identify analytics that can detect when the exception is taking place.

Subsequent to this analysis, we developed a purpose-built online system to collect feedback from the CATALYST technical and evaluation partners to prioritize which analytics get implemented:

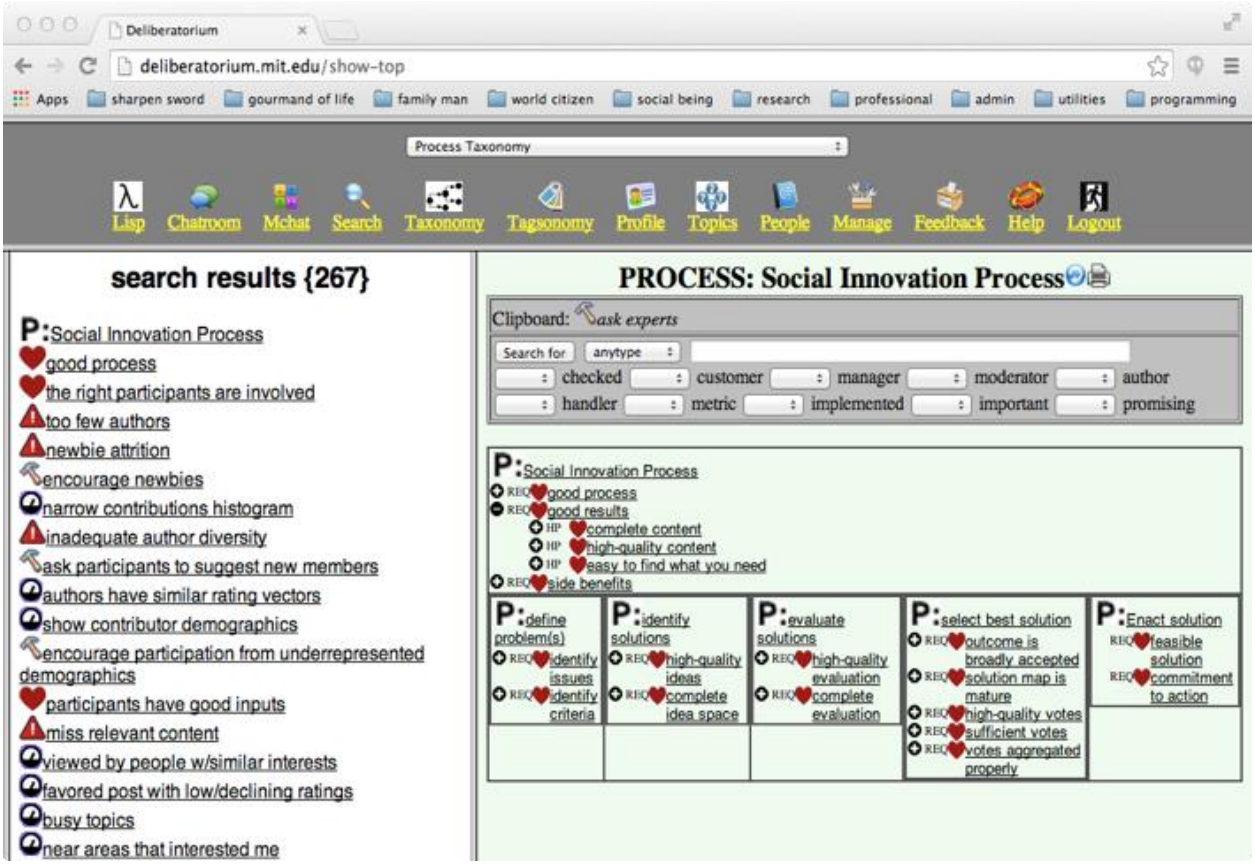


Figure 5 - Alert selection system

The CATALYST project has received funding from the European Union's Seventh Framework Programme for research, technological development and demonstration under grant agreement n°6611188

The results of this prioritization exercise guided subsequent development of the analytics server metrics and alerts library.

1.4 Visual analytics based on attention metrics

The CI analytics visualisations developed during CATALYST provide visual representations of conversations. Each visualisation has an analytical focus and the visualisations provide means to analyse a specific facet of a conversation. The CI Dashboard (https://cidashboard.net) contains all visualisations developed within the CATALYST project. Deliverable D3.9 (Liddo & Bachler, 2014) described the dashboard architecture and the first set of visualisations. Deliverable D4.6 (Ullmann, Liddo, & Bachler, 2014) reports on several additional visualisations and their evaluation. The dashboard has been extended with several visualisations since then.

The CI Dashboard makes use of the CI metrics server. Deliverable D3.5 (Klein, 2014) describes metrics provided by the server. These metrics have been extended since then, as described above. The CI Dashboard provides visualisations for these metrics.

The following sections highlight several visualisations of the CI Dashboard that make use of CI metrics delivered by the metrics server. Each section describes a particular metric and the chosen corresponding visualisation. It outlines the design and the rationale of each visualisation.

1.4.1 Activity bias visualisation

The activity bias visualisation introduced in deliverable D4.6 (Ullmann, Liddo, & Bachler, 2014) shows contributions plotted on a xy-plot (scatter plot). The visualisation aims to make visible the existence of biased activity patterns within conversation. A bias may exists if a conversation exhibits several distinct activity patterns. Figure 6 shows the activity bias visualisation. Each dot represents one contribution or several contributions in case they lie on the same coordinates.

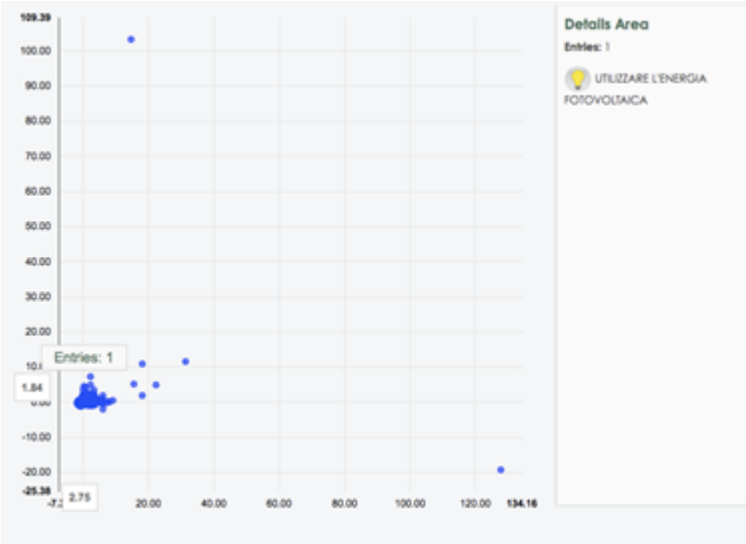


Figure 6 - Attention bias visualisation

The position of a contribution within the plot is determined by a metric of the UZH metrics server (see deliverable D3.5 (Klein, 2014) for more information about the metrics server). The underlying metric of this visualisation is based on singular value decomposition (SVD) (Golub & Kahan, 1965). Singular value decomposition is amongst others a popular technique for dimension reduction. For this visualisation the metric server calculates the singular value decomposition by taking as input contributions and their activity (e.g. viewing, editing, updating) and returns coordinates for each contribution. The visualisation shows the first two dimensions of the n-dimensional space returned by the metric server, which are also the most important ones. A scatter plot visualisation is a common way to represent the results of a SVD.

Clusters or groupings of conversations shown in the visualisation represent conversations that are similar according to the singular value decomposition. Their coordinates are close to each other. A visualisation showing several clusters represent a situation where contributions are similar within a particular clusters but different from other clusters. Each cluster can be seen as a

The CATALYST project has received funding from the European Union's Seventh Framework Programme for research, technological development and demonstration under grant agreement n°6611188

representation of a distinct pattern. If the visualisation shows more than one cluster than the used contribution data contain several different patterns. This may indicate that the contributions are biased regarding their activity pattern, meaning that there exists a set of contribution clusters that all have a different activity pattern. The activity bias visualisation aims at helping to determine if there are different patters in the underlying data. SVD, however, does not specify what the pattern is. The main function of this visualisation is to provide means to reveal the existence of patterns of conversations. Being aware of contributions having different activity patterns is important as it may lead to further exploration of the data in order to determine reasons of contributions having different activity patterns.

1.4.2 Rating bias visualisation

Similar to the activity bias visualisation, the rating bias visualisation is based on SVD. The metric takes into account all contributions and their ratings instead of contribution activity as for the activity bias visualisation. As the activity and rating bias visualisation are similar, the description of the underlying SVD can be found in the section about the activity bias visualisation. The visualisation represents the first two dimensions returned by the metric in a scatterplot visualisation. Figure 7 shows the rating bias visualisation.

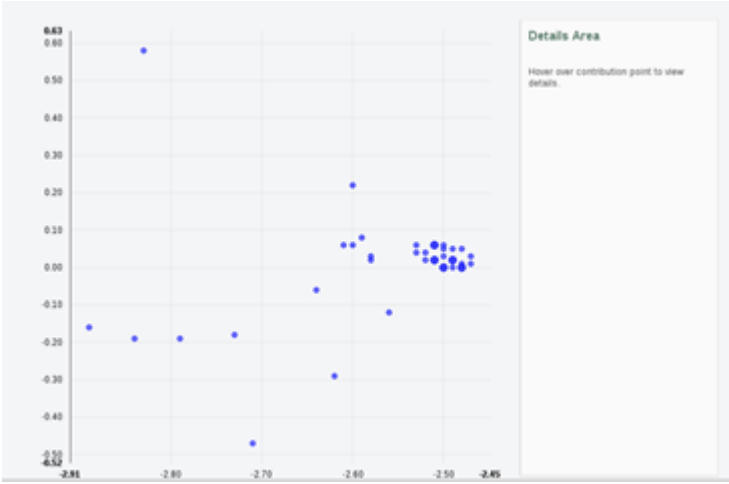


Figure 7 - Rating bias visualisation

Each dot represents a contribution. The coordinates of each dot are determined by the metric. A cluster indicates that the contributions within the cluster had a similar rating pattern. If the visualisation shows several clusters than this is an indicator that not all contributions were rated in a similar way. The rating behaviour of the users might have been biased as they rated contributions with different rating patterns.

1.4.3 Attention map visualisation

The attention map visualisation shows how equal or unequal attention is given to conversation threads of a conversation topic. The attention map visualisation shows an entire conversation as nested circles of contributions. The colouring of the circles indicate the equality or inequality of contributions by analysing the sub-parts of which a contribution is made of. The attention map visualisation uses the same visual principles of the 'conversation nesting visualisation' described in Deliverable D4.6 (Ullmann, Liddo, & Bachler, 2014). A conversation is understood as a hierarchy of contributions types, for example a group consists of issues, an issue consists of ideas, and idea can have supporting or counter arguments). A circle represents one of these contribution types, circles inside the contribution represent the sub-types of this contribution. Figure 8 shows an example of the attention map visualisation.

The CATALYST project has received funding from the European Union's Seventh Framework Programme for research, technological development and demonstration under grant agreement n°6611188

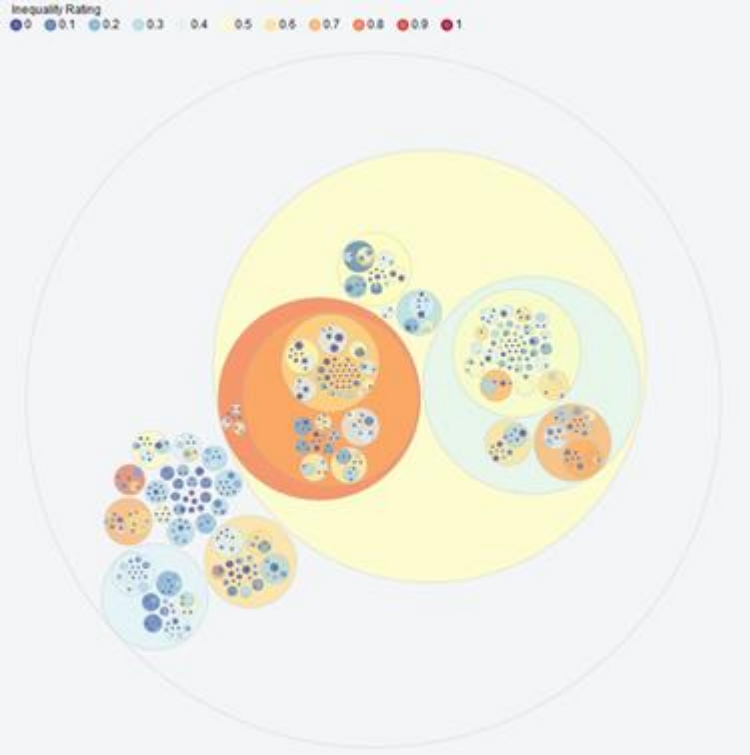


Figure 8 - Attention map visualisation

The colour of the circles is determined by a metric of the UZH metric server, which calculates the equality or inequality of a contribution based on its sub-contributions. The colour spectrum ranges from blue to red. Circles representing contributions in the blue spectrum are generally equally supported, while circles in the red spectrum represent contributions that are unequally supported. The metric made available by the metric server is based on the Gini coefficient. The metric server returns for each contribution a value between 0 and 1 (from equal to unequal), which is mapped to the colour range of the circles of the visualisation. As the visualisation shows the whole conversation in one picture, the visualisation makes it easy to spot areas that are equally supported or less equally supported. The interaction with the visualisation allows zooming into each circle allowing the close inspection of the contributions within a contribution.

1.4.4 Community interest network visualisation

The community interest network visualisation shows the interest of community participants in form of a network visualisation. The visualisation shows the whole conversation in form a network graph. Each node represents a contribution and the link between the nodes expresses that both note are in a relation to each other. For example, a node can be an issue and another node can be an idea. If the idea is part of the issue, then a link is drawn between both nodes. Figure 9 shows an example of the community interest network visualisation.

The CATALYST project has received funding from the European Union's Seventh Framework Programme for research, technological development and demonstration under grant agreement n°6611188

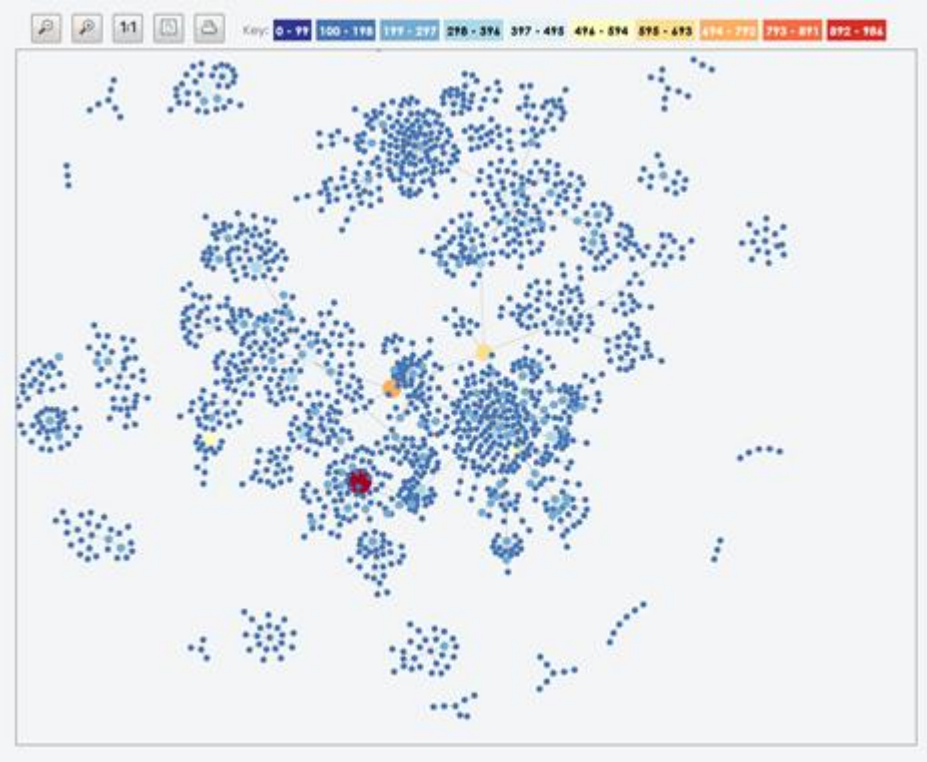


Figure 9 - Community interest visualisation

The interest of the community is depicted by the colouring of the nodes and the size of the nodes. The more red and large the node, the more interest received by the particular contribution represented by the node. The more blue and small the node, the less interest received by the contribution.

1.4.5 Sub-communities network visualisation

Similar to the community interest network visualisation, the sub-communities network visualisation shows a whole conversation in form of a network. Nodes are represented as coloured shapes. Figure 10 shows an example of the sub-communities network visualisation.

The CATALYST project has received funding from the European Union's Seventh Framework Programme for research, technological development and demonstration under grant agreement n°6611188

Figure 10 - Sub-communities network visualisation

Shapes that are the same represent similar contributions. The similarity is determined by a metric of the metrics server. Given a conversation, the metric returns clusters of contributions that tend to be looked at together. This metric uses singular value decomposition to find contributions whose activity is correlated. So, for example, we may find people who pay a lot of attention to the posts on stock markets also typically pay a lot of attention to the posts on banking. Posts with highly correlated activity can then be said to represent a 'theme' or 'topic' family. In the example above, the theme family would be something like 'finance'. So, this metric looks for topics, and tells you all the posts in that topic.
This approach can be also used to give recommendations (e.g. look at the other contributions in a cluster if you looked at one of the posts) as well as to reveal dependencies across issues in a map (i.e. if contributions from different issue branches have correlated activity). Note that those attention-based topic families are distinct from the semantic clustering described later.

1.5 Alerts and user feedback interface in LiteMap and DebateHub

The whole list of alerts implemented from the analytics in the Catalyst project can be found in the CI Dashboard website at: https://cidashboard.net/#alert (See Figure below).

The CI Dashboard interface consists of three main components:

1. A top part where the data is provided to the CI dashboard service (basically URLs of CIF formatted data of the debate to which Users want to apply the Alerts);
2. A central right part, where users can select the type of alerts they want to be calculated by the CI dashboard service;
3. A central left part where an example/preview of the Alerts widget is shown together with a list of information about dependency of each alert form specific data entry to be provided.

The CATALYST project has received funding from the European Union's Seventh Framework Programme for research, technological development and demonstration under grant agreement n°6611188

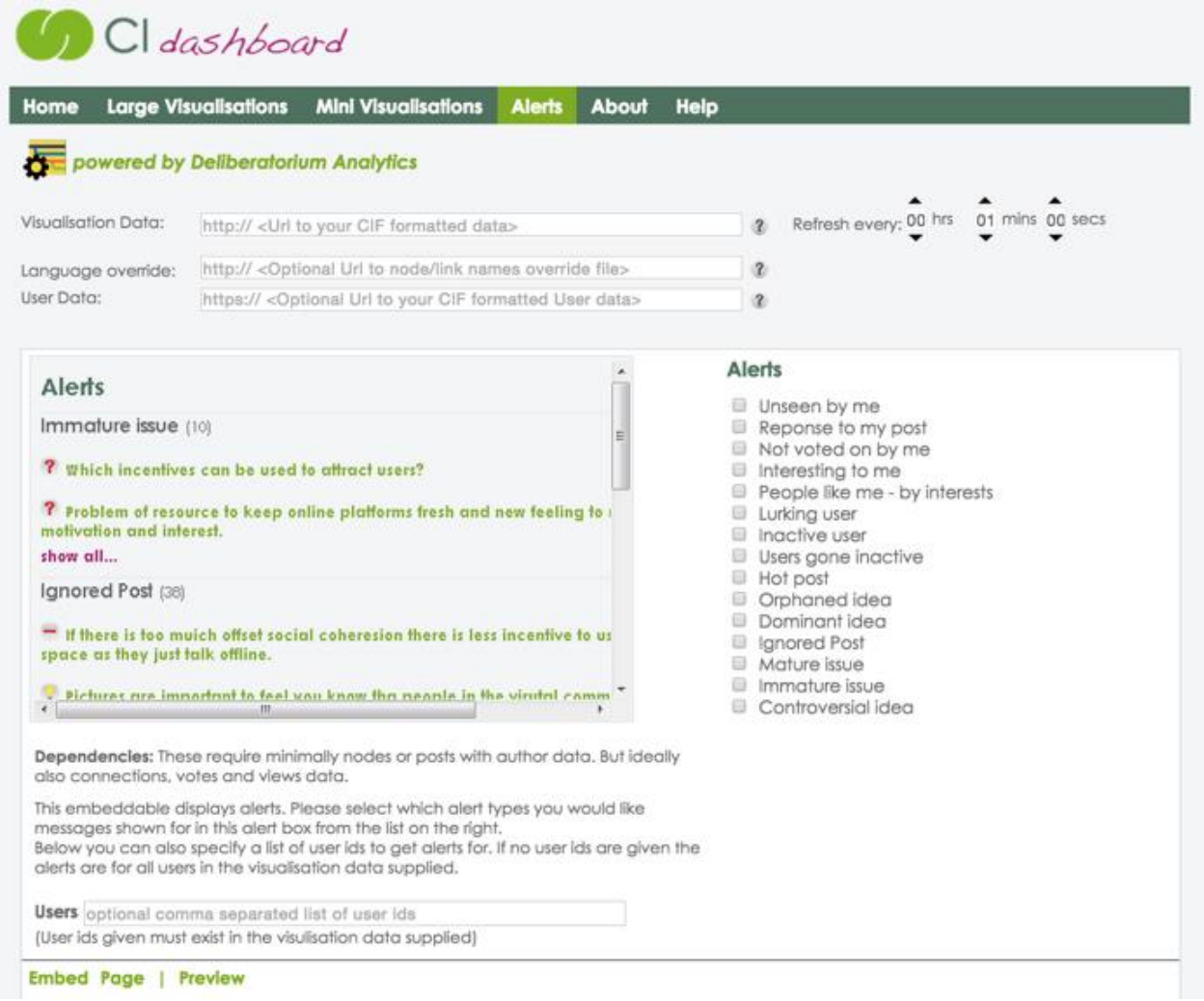


Figure 11 - CI Dashboard alerts interface

These alerts have been added to DebateHub and LiteMap in different clusters and with two different interfaces.

1.5.1 Alerts in LiteMap

Alerts can be classified in several ways besides the thematic distinctions above. In LiteMap we distinguish alerts between **Personalized Alerts** which mainly consist of alerts that are tailored to the logged in users and take into consideration their contributions and activities; and **General Alerts,** which are visible by all users at any point and do not require them to be logged in. These alerts are based on the overall debate data (contributions, users' activities, debate structure etc.).
The General Alerts are the following (taken from the list of alerts above): ignored post; mature issue; immature issue; hot post; orphaned idea; emerging winner; contentious issue; controversial idea.

The Personalized Alerts are extra alerts that logged-in users can see, and are built on logged in user's personal data. They are: unseen by me; response to my post; interesting to me.

1.5.1.1 LiteMap's alert interface

The CATALYST project has received funding from the European Union's Seventh Framework Programme for research, technological development and demonstration under grant agreement n°6611188

In Litemap Alerts are provided in the Alert sidebar which can be activated or deactivated on demand, and is displayed on the right side of the argument map. “A thumb up” icon was added to the left of each alert for users to provide quick feedback on the usefulness of the alerts.

Figure 12 - LiteMap alerts interface

1.5.1.2 User feedback Interface

By rolling over the thumb up icon a pop up message appears which says “Click if you found this alert useful”. If the user decides to click, a fading message is displayed which says: “Thanks for your feedback”.

1.5.2 Alerts in DebateHub

In DebateHub there are not general alerts, which is to say that no alerts are displayed to users who are not logged in. Debatehub has special moderator features available only to Debates’ Administrators. Therefore, for logged in users, alerts are divided into **Moderator Alerts** and **Personalized alerts**.

Moderator’s alerts are alerts built on general debate data, and which can help moderators to better manage the debate: for example by pointing users’ attention on immature or orphaned ideas, attracting lurkers to contribute, managing contentious issues etc.
Moderators can act upon the alerts suggestion by leaving “moderator’s comments” or by merging or splitting contributions.

Moderator Alerts are the following: Lurking user; ignored post; mature issue; immature issue; hot post; orphaned idea; emerging winner; contentious issue; controversial idea.

Personalized Alerts are alerts that logged-in users can see, and are built on their personal data. They show user-centric data: where their interests lay, attracting their attention to things they may have not seen or may be interested to see, and pointing out which

The CATALYST project has received funding from the European Union’s Seventh Framework Programme for research, technological development and demonstration under grant agreement n°6611188

contributions have been made by the community in response to their contribution. They are the following: unseen by me; response to my post; unrated by me; interesting to me.

Figure 13 - DebateHub alerts interface

In DebateHub **Moderator alerts** are displayed on the top right just below the moderators pictures list. Additionally, logged-in users (both debate participants and moderators) can see their alerts displayed at the bottom of the right sidebar under the **My Alerts bar** (See image below).

When clicking on one alert, the page scrolls down to the contribution which the alerts is pointing at, and the contribution is highlighted in yellow for a couple of seconds before fading away.

The CATALYST project has received funding from the European Union's Seventh Framework Programme for research, technological development and demonstration under grant agreement n°6611188

1.6 Semantic clusters

Whereas LiteMap and DebateHub focus on structured conversation, Assembl focuses on progressive structuration through harvesting. For that reason, deliberation analytics using deliberation structure are less relevant, and the focus was put instead on metrics useful to harvesters. I4P and UZH developed an analytics that detects clusters of related posts, based on latent semantic indexing[2], with basic linguistic pre-treatment[3]. We first experimented with the DBScan clustering algorithm[4] on the resulting semantic data, but later replaced it with Optics[5] clustering (Ankerst et al. 1999).

This development of those analytics was done as part of the community test with the 7th Sustainable Summer School. The goal was to give harvesters access to the semantic neighbourhood of a given message, allowed them to understand which messages of the discussion were covering similar topics, and to use cluster detection to propose grouping related messages under a topic.

2. Community testing of visual analytics

Several of the now existing visualisation that can be found on the CI Dashboard (http://cidashboard.net) have been evaluated in deliverable D4.6 (Ullmann, Liddo, & Bachler, 2014), either in the context of a field study examining the participants of the 'Design community 2014' or of lab study. In total 10 visualisations have been evaluated regarding their usefulness and usability as well as regarding users' task performance.
The detailed description of each of the visualisation, the evaluation design, and evaluation results can be found in D4.6. The following paragraphs summarise some of these experiences made during the evaluation.
The evaluation presented here is based on two studies. The first study was conducted as a field experiment in the open, why the second experiment took place in a usability lab. The participants of the field experiment were group members of the DebateHub[6] 'Design community 2014' group[7]. This group used DebateHub to discuss group issues, to come up with ideas and to weight the ideas with supporting or counter arguments. The group members have been invited to participate in the evaluation via Email. Participation was voluntarily.
The lab experiment participants voluntarily agreed to take part in the evaluation study. They followed the advertisement of the evaluation study distributed through several channels of the Open University.
The general setup for both studies was the same, while the concrete implementation differed in order to adapt to the context of the study. The building blocks of the general setup consisted of a background questionnaire, an introduction to the scenario, a phase where the participants explored the visualisation on their own time, a task, and a questionnaire to evaluate the usability of the visualisation.
The participants of the field experiment received an email with links to a questionnaire[8] for each visualisation. The questionnaire guided the participants through all building blocks of the evaluation.
The facilitator of the study guided the lab participants. They were asked to fill out the background questionnaire, they were verbally introduced to the scenario, had time to explore the visualisation on their own, they were asked the same task questions, and they had to fill out the same usability questionnaire.

[2] Using gensim, https://radimrehurek.com/gensim/
[3] Multi-lingual stemming with snowball: http://snowball.tartarus.org/, phrase detection and Tfidf normalization with gensim.
[4] Using the implementation from scikit-learn (Pedregosa et al., 2011)
[5] Our implementation, based on work by Brian Clowers: http://www.chemometria.us.edu.pl/download/optics.py
[6] http://debatehub.net/
[7] Data Available at: http://debatehub.net/group.php?groupid=98113864405029350014098719 56
[8] Created with the maQ-online questionnaire generator. Developed by Ullmann, T. D. (2004). maQ-Fragebogengenerator. Make a Questionnaire. Available online at: http://maq-online.de

The CATALYST project has received funding from the European Union's Seventh Framework Programme for research, technological development and demonstration under grant agreement n°6611188

The scenario consisted of asking the participants to imagine that they are tasked to make sense of large online conversations. These conversations are large enough that manual inspection of the contributions is neither reliable nor effective. Instead participants should use the analytics visualisations to make sense of the conversation.
Each visualisation contained a small task. Participants had to answer three to four questions for each visualisation. The participants could find all solutions to the task by using the visualisation. For example, two of the four questions for the user activity analysis visualisation were: 'How many times did the most active user contribute to the debate?' and 'How many counter arguments have been made in the whole debate?'
The data for the visualisations were taken from the CIF file generated from the conversation of the 'Design community 2014' group. The data were the same for both groups.

2.1 Evaluation

The evaluation presents the background information of the participants, their task performance, and the usability scores for each visualisation. The field experiment participants as well as the lab participants could stop the task anytime. In the case of the field experiments, the participants were not required to fill out all questionnaires. During the lab study on average two visualisations were evaluated per participant.

2.1.1 Background information

Field experiment: On average 7.4, mostly female, participants filled out the questionnaire for each visualisation. Most of them visited DebateHub between two to ten times. Most participants made one contribution, and a few made more than 10 contributions. Analytics dashboards were mostly new to them; they had slightly more familiarity with visualisations to explore data.
Lab experiment: 12 participants evaluated the visualisations (5 female and 7 male). Each visualisation was rated by five of them. Their familiarity with analytics dashboards ranged from novice to advanced, they were mostly novices with visualisations to explore data, but also all levels of familiarity (from novice to expert) were present.

2.1.2 Task performance

Table 1: Task performance shows the performance of the participants on the tasks for each visualisation. It shows the number of participants (N), the number of questions (Questions), and the percentage of correct answers. For example, the field experiment group answered in 90% of the cases correctly. Out of 40 answers 4 were answered incorrectly. The lowest amount of correct answers had the lab group for the debate network visualisation. 10 out of 15 answers were correctly answered.

Table 3. Task performance

		Field		Lab	
Visualisation	**Questions**	**N**	**Per cent**	**N**	**Per cent**
Quick overview	4	10	90	5	100
Debate network vis.	3	6	72	5	67
Conversation nesting	3	7	100	5	87
Activity analysis	4	4	75	5	80
User activity analysis	4	6	75	5	90

2.1.3 Usability

The usability of the visualisations was measured with the SUS usability questionnaire (Bangor et al. 2008, 2009; Brooke 2013). Table 2: Usability shows the results of the usability questionnaire. The table shows the calculated SUS usability indices and the average of the ratings to the question 'Overall, I would rate the user-friendliness of this visualisation as [worst, awful, poor, ok, good, excellent, best]'. The usability of all visualisations has been rated between ok and excellent. The participants of the field experiment

The CATALYST project has received funding from the European Union's Seventh Framework Programme for research, technological development and demonstration under grant agreement n°6611188

rated most visualisations as good, with the exception of the quick overview visualisation, which was rated as ok. Similarly was the lab group, which rated all but the debate network visualisation as good.

Table 4. Usability

	Field experiment			Lab experiment		
Visualisation	**N**	**SUS**	**Over.**	**N**	**SUS**	**Over.**
Quick overview	9	57.50	4.22	5	86.0	5.2
Debate network vis.	6	67.08	4.50	5	68.0	4.4
Conversation nesting	6	81.67	5.83	5	78.5	5.4
Activity analysis	4	53.75	4.50	5	79.5	5.4
User activity analysis	6	67.08	4.67	5	71.0	5.2

2.1.4 Discussion

Overall, the participants performed well on the task independently from the two testing conditions (in the wild and in the lab). This is an encouraging finding, considering that most of the participants were novices regarding analytics dashboards and visualisations to explore data. Task performance differed. Both groups performed very well on the task of the quick overview visualisation, and both groups made mistakes in less or equal to one third of the questions of the debate network visualisation. More or equal to 75% of the questions got answered correctly for all other visualisations.
Most visualisations were rated as having a good usability. Overall, the lab participants rated the usability as higher than the field experiment group. Differences have been found in the quick overview visualisation and the activity analysis visualisation, which was found more usable in the lab group. All other three visualisations have been on the same level. The participants had more problems with answering the tasks for the debate network visualisation. The low ratings, relative to the other ratings, of the usability may indicate usability issues that need to be followed up.

Unfortunately, this testing was done in an early phase of the project, and we did not get to test the more complex visualizations and metrics. Some of the visualizations, such as activity bias and rating bias, depend on somewhat complex calculations, and we have found it difficult to convey to users what the analytics mean in terms of the discussion. Without formal testing, users who have seen those visualizations found them difficult to understand. Attempts to explain the underlying principles were not too successful, but on the other hand, users were also perplexed by a naked score extracted from the metric without the context of where the score came from. We now believe that the best solution is to translate the metric's signal back into the basic entities of the deliberation; and we are working on some visualizations based on this (much more demanding) approach, one example of which is presented below.

3. Community testing of alerts

By the time the alerts were ready for testing, we had two community tests scheduled: Loomio and the Seventh Sustainable Summer School.

3.1 Loomio test of harvester alerts

The Loomio testing was designed to test the use of Alerts, built from deliberation analytics, in the context of the LiteMap tool.
A series of meetings were first organized to figure out what type of discussion topics and testing structure would fit both research and community interest.

Research questions:
Does the use of Litemap improve and stimulate existing online discussions?

The CATALYST project has received funding from the European Union's Seventh Framework Programme for research, technological development and demonstration under grant agreement n°6611188

Does the alert functionality of Litemap stimulate more/better mapping activity?

3.1.1 Experimental design

In the end we agreed with Loomio on the following AB testing structure:

34 Litemap participants were split into two groups - A and B;
Two separate Litemap instances for A and B were created with alert and alert-less functionality respectively.
A Loomio group was created where participants could ask questions about the project and give feedback on Litemap as the study progresses. Instructions and outline of the test was provided to this group.
Both groups were also emailed the instructions in the form of a link to the video demo How to Harvest Online Content with Litemap[9], and were invited to create a Litemap account, and join the Loomio group.
Group A was invited to harvest content on the theme of "Business models for open-source social impact ventures", and provided a link to the Long-term financial sustainability of Loomio[10] discussion as an example of content and other resources elsewhere (forthcoming).
Group B was invited to harvest content on the theme of "Role of online democracy tools in government". A link was provided to the Political use of Loomio[11] discussion as an example of content and other resources elsewhere.
After that, there was to be a two week phase where comments would be posted in each of these discussions, informing participants about the mapping of their discussion (and the general topic), and inviting existing discussants to participate in the harvesting by joining the study's Loomio group and following the instructions (above).
Then the harvesting phase would begin. When the map would reach a certain size threshold, or a specified period of time (two weeks) would pass, a link to the map would be posted in each of Loomio discussions, and the discussants would be asked to rate its usefulness.
Participants would be encouraged to provide feedback on the tool and study as it unfolds, allowing us to make minor adjustments and clarify as necessary.
After the mapping activity would end, or a specified time period would have passed, we would follow up with harvesters and discussants through surveys about their experience.

From the comparison of performance and users' feedback from group A (with alerts) and B (without alerts), we'd be able to establish if the presence of alerts made a difference in the harvesting work. At the same time the answers to the Litemap survey sent to Loomio discussion participants would have provided insights on if/and to what extent LiteMap had improved Loomio's discussion.

3.1.2 Testbed outcomes

Unfortunately the test failed to attract participants, only 4 people of the 34 recruited by Loomio participated to the discussion in the two groups, and only two people created a LiteMap account and contributing only a couple of ideas to the maps.

While exploring the reasons for this failure to engage the community we learned that **Loomio is a platform used worldwide but does not host large-scale debates**.
Loomio's moderator Simon Tegg explained to us that, even in the most successful debates, there is usually a medium of 5-9 participants per conversation. This was not what we had envisioned and what we intended in the Loomio's open call submission.

As a lesson learned for the future, we now know that when recruiting for testing organizations which bring online discussion tools, we should not simply ask communality partners: 'How many users do you have?' Or 'What is the geographical scale of your outreach?'

[9] https://www.youtube.com/watch?v=uBzVLUDIxE8
[10] https://www.loomio.org/d/mUOKA6vc/long-term-financial-sustainability-for-loomio
[11] https://www.loomio.org/d/IGSO7guf?page=1

The CATALYST project has received funding from the European Union's Seventh Framework Programme for research, technological development and demonstration under grant agreement n°6611188

Many communities in fact have very wide mailing lists from people all across the globe but are unable to mobilize large numbers of people at once in one discussion. Basically the large-scale users list is simply constituted of widely fragmented small group of people, which do not represent a community and cannot support large-scale discussions.
That is why the question to ask to recruit people should rather be: 'What is the medium number of participants you are able to engage in a single debate on your platform'? **In Catalyst we have found out that few community partners can meet this requirement.**

Another minor problem was worldwide participation. Website access was slow from New Zealand and Australia. Several options have been tried to improve the speed but possible solutions did not depend on LiteMap improvements so in the end we had to cut out participants from that area of the world. In the end we had 34 people out of the initial 47 recruited from Loomio.
In any case this caused only a loss of a few participants, while the main issue remained that the 34 recruited people did not even sign up to the website.

We therefore explored a way around solution, which allowed us to **carry on the interoperability testing of Loomio with other 3 Catalyst tools** in the following user scenario:

We pick an existing Loomio discussion, which has received enough participation.
The Catalyst's tech team uses Assembl's Loomio feed import to get Loomio's posts into Assembl and then use Assembl to mark up and structure key ideas out of the existing Loomio discussion in form of an ideas' thread
Then one harvester works to IBISify the discussion, that is to say, she creates an argument map out of the Assemble ideas' thread.
Then Loomio embed the argument map and other relevant discussion's visual analytics from the CI Dashboard in the Loomio discussion page.
Finally, ask feedback on the discussion's visualisations and analytics to the Loomio community with a very small survey we sent to them by email. (the survey should aim to answer questions such as: Do they find these visualisations useful, understandable, engaging etc?)

Loomio's was unable to support stages 3 and 4 so the final interoperability test was conducted by the Catalyst tech team and has been showcased in this demo movie: https://www.dropbox.com/s/rivf95cnaz50gs7/LiteMap-Loomio-Assembl-CIDashboard-Interoperability_Demo.mov?dl=0

As a consequence of these various shortcomings Loomio's second grant payment was not granted.

3.2 Seventh Sustainable Summer School

This test is described in detail in D4.3. Unlike the Loomio test, the summer school was active, but on a very small scale. For that reason, many alerts that would have been useful in a larger conversation were of limited use in this context, and we had to design new alerts that could be used in the context of the conversation.

3.2.1 Participation-centric analytics

In particular, participation-centric alerts designed to alert moderators to participant activity (e.g. the alert that a user has gone inactive) are of limited use to moderators in a very small-scale debate such as that in the school, because moderators are familiar with individual participants and can notice activity patterns directly.
Nonetheless, the basic statistics about user activity presented by Assembl were consulted by the moderators, and moderators did say that they would appreciate the more detailed participation analytics and alerts in the context of a larger conversation.
There was also a perception that, even if moderators were made aware of users who were not actively participating, it was very difficult to use that information to enjoin them to participate without annoying them.

The CATALYST project has received funding from the European Union's Seventh Framework Programme for research, technological development and demonstration under grant agreement n°6611188

For similar scale reason, the alerts telling a participant about other participants with similar activity patterns, or ideas interesting to such similar participants, were seen as potentially of very limited use in a context where participants were familiar with one another and were not pursued.

3.2.2 Semantic clusters and semantic proximity

For all those reasons, we developed the clustering metrics described above, under the assumptions that they could be more useful to harvesters in this context, with a good number of posts by very few users.

This development was done during the discussion, and was thus only made available near the end of the process. Nonetheless, the late discussion was also when the messages were the most numerous, and when the feature was most needed. Data on semantic neighbourhoods did help harvesters to find related messages elsewhere in the discussion. This was especially true given the structure of the discussion, where the overall discussion was divided in three broad phases: first challenges, then solutions, then sustainable design. The same broad themes were discussed within each phase, but the grouping of the conversation in phases made it particularly useful to examine each theme across the phases through semantic neighbourhoods. While the final synthesis was structured in phases, the content streams of each theme were also presented during a live meeting with the students, and the semantic neighbourhood function allowed to extract those content streams more easily and with better recall.

As for the clustering, the first version of the clustering algorithm (with DBScan) did reveal a few subsets of posts that were not reflected in the table of ideas; notably, a subdivision of a discussion around waste at work into paper and electricity waste. However, those posts had been grouped deliberately under a broader heading, and the subdivision, though meaningful, was not deemed essential to the structure of the table of ideas. The optics-based algorithm also identified clusters whose content was mostly contained within an idea, but with a few outlying posts that should probably have been classified with that idea but were not. There is little doubt that this would have allowed more exhaustive harvesting. In other cases, some of clusters correspond to concepts that were already harvested in ideas, and we can surmise that it would have helped the harvesters if the cluster had been found before the idea had been identified. However, because those clustering analytics came quite late, we do not have live examples where the cluster preceded the harvester's action.

Overall, our limited community testing shows a valuable use for semantic neighbourhood search, and our first experiments with automatic clustering show that it can exhibit some meaningful post clusters most of the time. It remains to be proven whether those clusters will actually assist harvesters in populating the table of ideas, but we have every reason to expect cases of this to emerge naturally as we make those features available to harvesters.

4. Post-hoc testing

Following the qualified success of one of the tests, and the failure of the second to gather useful data, we decided to do more testing of the analytics in laboratory condition. Our aim was to demonstrate that the metrics could detect useful conditions in real discussion data, even if we could not get a significant community to use those alerts at this stage in the process. We evaluated both the alerts and metrics components of the CATALYST analytics server.

4.1 Alerts

The alerts were evaluated by assessing how often clearly dysfunctional patterns (e.g. a user rates a post without viewing the underlying arguments) occur in real-world deliberations. Clearly, if it turns out that such dysfunctional patterns appear very rarely in representative deliberation maps, then the alert is unlikely to add much value to a deliberation. To test this, we selected three representative alerts that seem to represent clearly dysfunctional patterns:

The CATALYST project has received funding from the European Union's Seventh Framework Programme for research, technological development and demonstration under grant agreement n°6611188

- Rating ignored argument
- Rating ignored competitor
- Unseen response

We then assessed how frequently these alerts appeared in an online engagement wherein 189 participants created a deliberation map with nearly 1600 issues, ideas and arguments (on the topic of bio-fuels use in Italy):

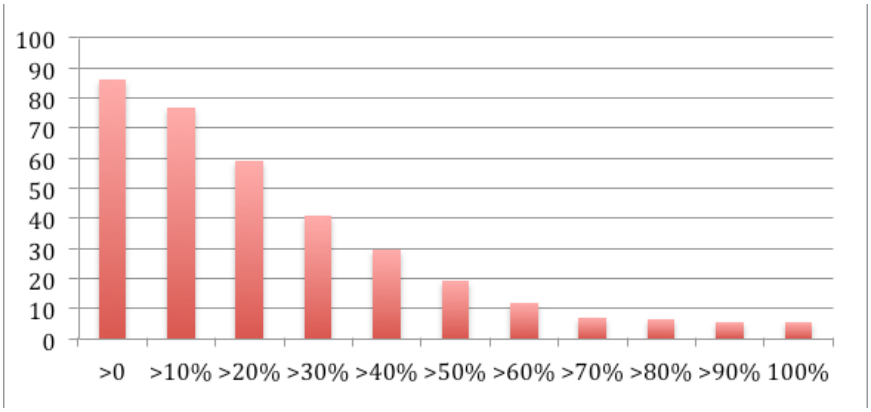


Figure 14 - How often users ignore arguments when rating posts

Roughly 86% of all participants rated at least some posts without viewing relevant arguments, and almost 6% of the participants rated posts without seeing *any* of the relevant arguments. On participants, users did not view roughly 37% of the relevant arguments when rating posts. This phenomenon could occur either because the participants ignored existing arguments when rating a post, or because the arguments were added *after* the user rated the post.

Remarkably, roughly 96% of the participants did not view at least some competing ideas (i.e. ideas responding to the same issue) when rating an idea. This is potentially important because it is usually important to assess the quality range of a set of ideas before assigning any of them a label such as "very promising" versus "somewhat promising".

We also found that participants, on average, did not view the responses to roughly 16% of the posts they authored, and that they miss roughly of these responses all told. This is potentially important because responses to their posts - e.g. rebuttals of their arguments, arguments against their ideas - would be likely to elicit further engagement in the deliberation process.

In all these cases, it seems clear that alerting users to view missed posts relevant to their contributions and ratings has the potential for significant impact on the deliberation.

4.2 Metrics and visualization

Metrics were evaluated by how assessing to what extent they make it possible to recognize important deliberation patterns that would be otherwise difficult to detect without such metrics e.g. if the user could only browse the deliberation map itself. For this purpose we chose to look at the metrics which allow users to assess to what extent the participant community's preferences concerning an issue are polarized into opposed camps, or not. We started with the following simple deliberation map, consisting of a single issue, two competing ideas, and a total of six arguments:

The CATALYST project has received funding from the European Union's Seventh Framework Programme for research, technological development and demonstration under grant agreement n°6611188

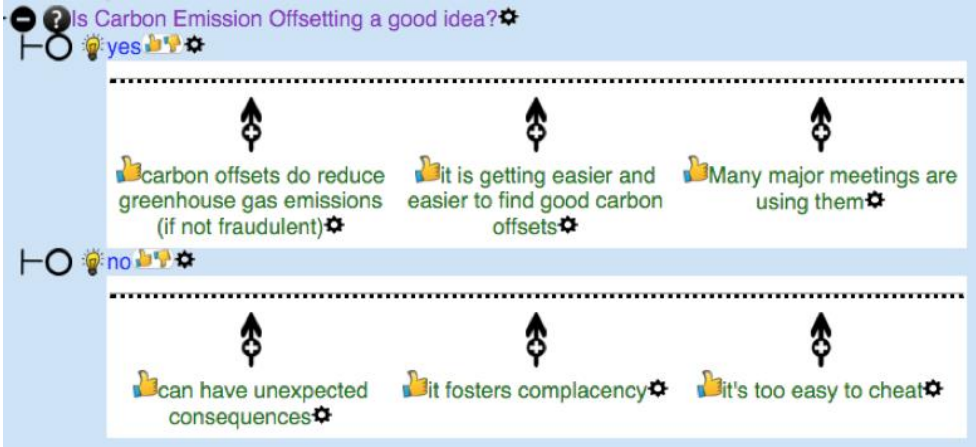


Figure 15 - Simple deliberation map

We compared two situations: one where eight simulated users provided uncorrelated (randomly selected) ratings for the posts in the map, and a second where the eight simulated users were divided into two balkanized groups, where one half of the users like all the posts that the other users do not, and vice versa. To keep the example simple, we assume that all users rate posts as either a "1" (strongly dislike) or a "5" (strongly like).

In the uncorrelated setting, the ratings from each user, in our example, are as follows:

Table 5. Uncorrelated ratings example

post	**u1**	5.1.	5.1.	5.1.	5.1.	5.1.	5.1.	5.1.
idea: yes	1	5	5.1.	5.1.	5.1.	5.1.	5.1.	5.1.
pro: carbon offsets do reduce greenhouse gas emissions (if not fraudulent)	1	5	5.1.	5.1.	5.1.	5.1.	5.1.	5.1.
pro: it is getting easier and easier to find good carbon offsets	5	1	5.1.	5.1.	5.1.	5.1.	5.1.	5.1.
pro: Many major meetings are using them	1	1	5.1.	5.1.	5.1.	5.1.	5.1.	5.1.
idea: no	5	5	5	5.1.	5.1.	5.1.	5.1.	5.1.

The CATALYST project has received funding from the European Union's Seventh Framework Programme for research, technological development and demonstration under grant agreement n°6611188

pro: can have unexpected consequences	1	5	5.1.	5.1.	5.1.	5.1.	5.1.	5.1.
pro: it fosters complacency	5	1	5.1.	5.1.	5.1.	5.1.	5.1.	5.1.
pro: it's too easy to cheat	5	5	5.1.	5.1.	5.1.	5.1.	5.1.	5.1.

In the fully balkanized setting, the ratings from each user, in our example, are as follows:

Table 6. Balkanized ratings example

post	**u1**	6.1.1	6.1.1	6.1.1	6.1.1	6.1.1	6.1.1	6.1.1
idea: yes	1	1	6.1.1	6.1.1	6.1.1	6.1.1	6.1.1	6.1.1
pro: carbon offsets do reduce greenhouse gas emissions (if not fraudulent)	1	1	6.1.1	6.1.1	6.1.1	6.1.1	6.1.1	6.1.1
pro: it is getting easier and easier to find good carbon offsets	1	1	6.1.1	6.1.1	6.1.1	6.1.1	6.1.1	6.1.1
pro: Many major meetings are using them	1	1	6.1.1	6.1.1	6.1.1	6.1.1	6.1.1	6.1.1
idea: no	5	5	6.1.1	6.1.1	6.1.1	6.1.1	6.1.1	6.1.1
pro: can have unexpected consequences	5	5	6.1.1	6.1.1	6.1.1	6.1.1	6.1.1	6.1.1
pro: it fosters complacency	5	5	6.1.1	6.1.1	6.1.1	6.1.1	6.1.1	6.1.1

The CATALYST project has received funding from the European Union's Seventh Framework Programme for research, technological development and demonstration under grant agreement n°6611188

pro: it's too easy to cheat	5	5	6.1.1	6.1.1	6.1.1	6.1.1	6.1.1	6.1.1

Note that the ratings histogram for all the posts in the uncorrelated and fully balkanized conditions would be identical, as follows:

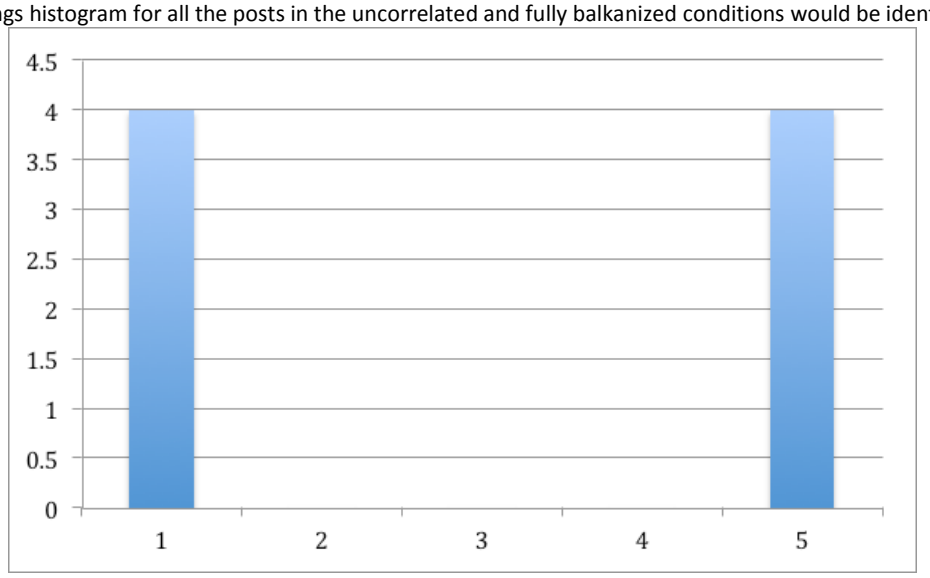


Figure 16 - Balkanized rating histogram

The existence and nature of any balkanization would be completely opaque to a user looking at the deliberation map, especially if (as is commonly done to encourage honest ratings), the ratings are anonymized.

The presence of absence of balkanization is, however, quickly revealed by simple visualizations of the appropriate metrics from CATALYST analytics server. If we plot the posts according to the first two (most predictive) eigenvectors returned by the interest space post coordinates metric, we can see that there is no clear clustering:

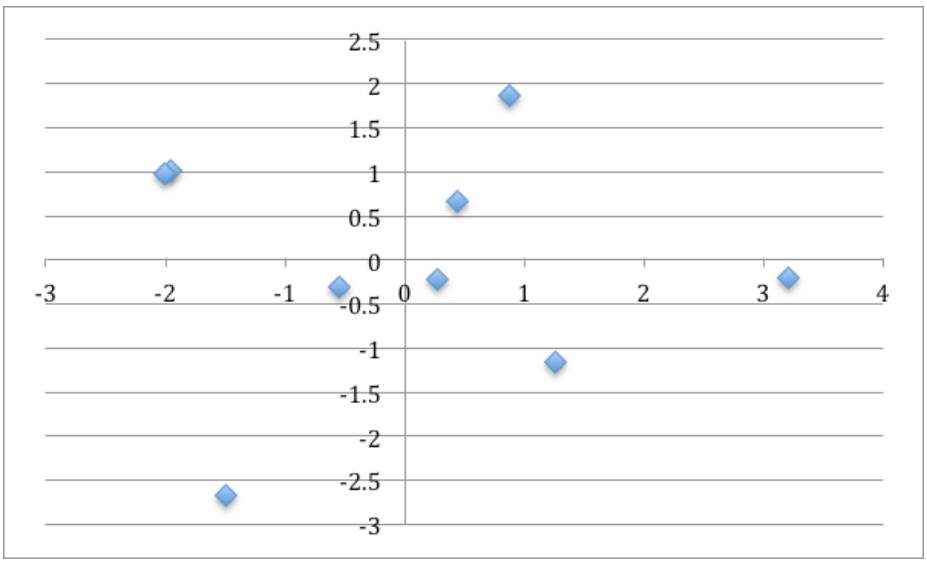


Figure 17 - Eigenvector decomposition of uncorrelated interest space

The CATALYST project has received funding from the European Union's Seventh Framework Programme for research, technological development and demonstration under grant agreement n°6611188

This is borne out by the fact that the support space clustering metric returns a value of 0 i.e. no clustering was observed.

By contrast, if we plot the posts from the fully balkanized case according to the first two (most predictive) eigenvectors returned by the interest space post coordinates metric, the clustering (and thus balkanization) becomes clear:

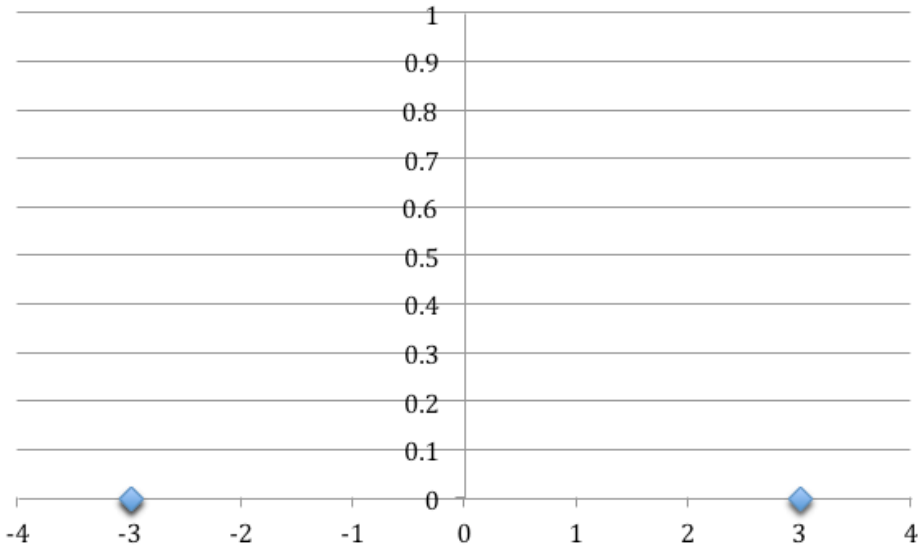


Figure 18 - Eigenvector decomposition of balkanized interest space

The clustering coefficient, in this simple example, is 1.0.

Finally, we can use the interest space dimensions metric to visualize the fault line across which the two balkanized groups are divided. The post weights for the most predictive eigenvector are as follows:

Table 7. Eigenvector analysis

idea: yes	0.35
pro: carbon offsets do reduce greenhouse gas emissions (if not fraudulent)	0.35
pro: it is getting easier and easier to find good carbon offsets	0.35
pro: Many major meetings are using them	0.35
idea: no	-0.35
pro: can have unexpected consequences	-0.35
pro: it fosters complacency	-0.35
pro: it's too easy to cheat	-0.35

If we simply color code the posts deliberation map according to the valence of the post weights (green for support, red for opposition), we can immediately see that the community is divided over whether carbon offsetting is a good idea or not:

The CATALYST project has received funding from the European Union's Seventh Framework Programme for research, technological development and demonstration under grant agreement n°6611188

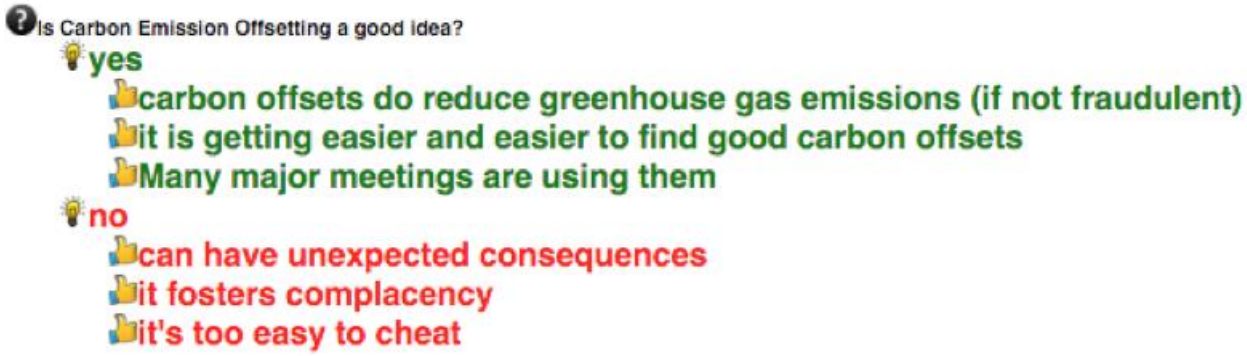


Figure 19 - Topic tree color-coded by balkanized community

This represents, we believe, a compelling demonstration of how metrics provided by the CATALYST Analytics Server can allow system users to derive a qualitatively deeper understanding of the "health" of the deliberation, and how mapping back abstract mathematical results to the original discussion entities can improve intelligibility.

4.3 Suggestions based on semantic clusters

The initial positive response of the design summer school moderators to semantic distance and clustering encouraged us to design further alerts that would go beyond clustering and provide direct help to harvesters. In bottom-up tools like Assembl in particular, harvesters are trying to extract semantic clusters by hand from harvested content, and this allows us to use the semantic proximity machinery to answer two very different kinds of questions:

Do the automatic clusters correspond to meaningful concepts? In which case:
Can the automatic clustering help enrich the harvester-defined content groups around ideas?
Can the automatic clustering help carve out new content groups within those already identified?
Do the content groups correspond to semantic clusters?

We have focused on the first kind of questions, due to their immediate practical use, but we have done some work of approaching the problem by the other end: It is possible to measure how much a given group of content (posts) defined by a harvester stands in contrast to neighbouring groups by calculating the silhouette score (Rousseeuw, 1987) of the group's content in the conversation. We had to adapt the algorithm slightly, as silhouette scores are a global measure of clustering, and we wanted to measure the score of a single group; also, the same message can be associated with a plurality of groups (ideas) by harvester, and the silhouette algorithm expects a single category per point. We have also measured what we call the internal silhouette score of a group, which is the silhouette score that corresponds to the subdivision of the group's content in its immediate sub-ideas. In general, we have found significantly positive scores (between 0.1 and 0.2, where zero should represent a random grouping of messages and negative numbers can represent a grouping that runs counter to computed clusters) in our conversations, at least in the leaf ideas. The ideas closer to the root of the thematic tree are usually too general to be distinguishable by semantic analysis, but the consistent positive results lower down in the tree would tend to confirm that the semantic analysis is not wholly irrelevant to the practice of harvesting.

So in theory, it should be possible for a recommender to propose adding unclassified content to the "nearest" few groupings, and to see whether the silhouette score is improved; and to subdivide each idea in many ways, and to see which subdivisions have high global silhouette scores.

In practice, the second algorithm is computationally intensive, and we have found it more practical to reuse the clusters from our earlier work (using the Optics algorithm) to accomplish both tasks, at least as an initial implementation. So our recommender takes each optics cluster, and finds the content groups with the most shared content. It then attempts the following:

The CATALYST project has received funding from the European Union's Seventh Framework Programme for research, technological development and demonstration under grant agreement n°6611188

- Adding the rest of the cluster to the group, to test whether that group's local silhouette score is improved;
- Using the cluster to introduce a new subgroup in that group, to see whether the group's internal silhouette score is improved;
- A combination of both strategies above.

In the case of additions, the recommender chooses the idea most improved for a given cluster, and in the case of partitions, it chooses the cluster that has the most impact on a given idea. Combinations have priority on either.

Though still partial, early results are encouraging. We have asked harvesters to evaluate clusters found by our algorithm in existing real conversations, and compare the usefulness rating of real clusters to fake clusters. The fake clusters are constructed by reverting the optics algorithm, and suggestions that minimize the silhouette score were presented to harvesters and mixed randomly with suggestions that improve it.

The results are somewhat mixed; many results from the suggestion engines were considered of limited use by harvesters, whereas many random results were found useful. Still, overall, engine recommendations were judged more useful than random. (Additions and partitions were analysed separately, because the former are scored by a delta to the outer silhouette score, whereas the latter are scored by a delta to the inner silhouette score. Mixed additions and partitions are analysed as partitions.)

Table 8. Cluster ratings by harvesters

Type	Suggestion or random	# useful	# useless
addition	suggestion	30	23
addition	random	6	17
partition	suggestion	19	21
partition	random	5	12

However, this report fails to take into account the fact that, within each category (suggestion or random), the score delta indicates degree of certainty of the result. Looking at the average score delta within each category, we obtain the following:

Table 9. Average silhouette score for cluster categories

Type	Judged useful	average score delta
addition	yes	0.029
addition	no	0.023
partition	yes	0.025
partition	no	0.008

Those results are not, however, significant, according to a Student's t-test:
additions: statistic=0.62, p value=0.54

The CATALYST project has received funding from the European Union's Seventh Framework Programme for research, technological development and demonstration under grant agreement n°6611188

partitions: statistic=1.14, p value=0.26

Nonetheless, the fact that so many of the additions or partition suggestions were judged useful by the harvesters is encouraging in its own right, and the cluster analysis will be included in the Assembl platform.

The CATALYST project has received funding from the European Union's Seventh Framework Programme for research, technological development and demonstration under grant agreement n°6611188

Conclusions and future work

While there has been substantial effort devoted to manually-coded, *post-hoc* metrics on the efficacy of on-line deliberations (Steenbergen et al., 2003) (Stromer-Galley, 2007) (Trénel, 2004) (Cappella et al., 2002) (Spatariu et al., 2004) (Nisbet, 2004), existing deliberation technologies have made only rudimentary use of automated *real-time* metrics to foster better emergent outcomes during the deliberations themselves. The core reason for this lack is that, in existing deliberation tools, the content takes the form of unstructured natural language text, limiting the possible deliberation metrics to the analysis of word frequency statistics, which is a poor proxy for the kind of semantic understanding that would be necessary to adequately assess deliberation quality. One of the important advantages of using argument maps to mediate deliberation is that they allow us, by virtue of their additional semantics, to automatically derive metrics that would require impractically resource-intensive manual coding for more conventional social media.

We have also shown how exception analysis can identify useful deliberation metrics, as well as how a collective intelligence approach can be used to gather community input on which metrics are most useful.

It seems clear overall that deliberation analytics can serve a useful role in medium to large-scale moderated conversations, whether through visualizations or alerts. However, one conclusion of our study is that, at least in the case of the more mathematically grounded alerts, their utilisation by community moderators requires designing a visualization that can relate those raw results to deliberation entities. For that reason, we will continue work on visualization. The CI dashboard allows us to test new visualizations comparatively easily, as we find new opportunities for large-scale conversations.

In particular, participation and interest metrics have been shown to be of limited use for smaller conversations, but we look forward to testing them in a larger-scale debate. Harvesters have declared an interest in semantic clustering alerts, and we will endeavour to refine those functions.

Though we think we have demonstrated the usefulness of many deliberation analytics, and shown the potential usefulness of many more, we were hoping to show that they could assist community managers to improve the quality of the conversation, and the limited scale of our tests has been such that un-assisted human intelligence could have achieved the same effect. So that research question remains open for now, though we have set in place a research apparatus that we hope could answer that question in the near future.

Beyond that, we were hoping that deliberation analytics could also help participants become aware of the community's collective communication dynamic, and raise the level of collective intelligence of the community. This question remains open, and our results show that the pedagogical element cannot be underestimated. In particular, past work by Mark Klein has shown that high-quality deliberative structure could be achieved in a moderated context (Klein 2012). It is yet unclear whether deliberation analytics could allow a comparable structure to emerge without moderation, through a combination of harvesting, just-in-time advice on (micro) deliberation structure, and alerts and visualization on macro conversation dynamics. Though we have shown the usefulness of deliberative analytics in assisting the harvesting process, and in identifying these conversation dynamics, measuring their impact on the capacity for self-organization of collective intelligence communities remains a question for future research.

The CATALYST project has received funding from the European Union's Seventh Framework Programme for research, technological development and demonstration under grant agreement n°6611188

References

Ankrest, M., Breunig,M., Kriegel, H. & Sander, J. (1999). OPTICS: Ordering Points To Identify the Clustering Structure, Proc. ACM SIGMOD'99 *Int. Conf. on Management of Data*, Philadelphia PA, pp. 49-60. http://doi.org/10.1145/304181.304187

Bangor, A., Kortum, P., and Miller, J. Determining what individual SUS scores mean: Adding an adjective rating scale. Journal of usability studies 4, 3 (2009), 114–123.

Bangor, A., Kortum, P.T., and Miller, J.T. An Empirical Evaluation of the System Usability Scale. International Journal of Human-Computer Interaction 24, 6 (2008), 574–594.

Brooke, J. SUS: a retrospective. Journal of Usability Studies 8, 2 (2013), 29–40.

Cappella, J. N., Price, V., & Nir, L. (2002). Argument Repertoire as a Reliable and Valid Measure of Opinion Quality: Electronic Dialogue During Campaign 2000. *Political Communication*, 19(1), 73 - 93.

Golub, G., & Kahan, W. (1965). Calculating the Singular Values and Pseudo-Inverse of a Matrix. *Journal of the Society for Industrial and Applied Mathematics Series B Numerical Analysis*, *2*(2), 205–224. http://doi.org/10.1137/0702016

Klein, M., & Bernstein, A. (2004). Towards High-Precision Service Retrieval. *IEEE Internet Computing Journal*, 8(1), 30-36.

Klein, M. (2012). Enabling Large-Scale Deliberation Using Attention-Mediation Metrics. *Computer-Supported Collaborative Work*, 21(4), 449-473.

Klein, M. (2014). *Deliberation analytics* (No. D3.5). University of Zurich. Retrieved from http://catalyst-fp7.eu/wp-content/uploads/2014/06/CATALYST_D3.5.pdf

Liddo, A. D., & Bachler, M. (2014). *Collective Intelligence Dashboard* (No. D3.9). Milton Keynes: The Open University. Retrieved from http://catalyst-fp7.eu/wp-content/uploads/2014/06/CATALYST_D3.9.pdf

Nisbet, D. (2004). Measuring the Quantity and Quality of Online Discussion Group Interaction. *Journal of eLiteracy*, 1122-139.

Pedregosa et al. (2011) Scikit-learn: Machine Learning in Python, JMLR 12: 2825-2830. url: http://jmlr.csail.mit.edu/papers/v12/pedregosa11a.html

Rousseeuw, P. (1987). Silhouettes: a Graphical Aid to the Interpretation and Validation of Cluster Analysis. *Computational and Applied Mathematics* 20: 53–65. doi:10.1016/0377-0427(87)90125-7.

Spatariu, A., Hartley, K., & Bendixen, L. D. (2004). Defining and Measuring Quality in Online Discussions. *The Journal of Interactive Online Learning*, 2(4),.

Steenbergen, M. R., Bachtiger, A., Sporndli, M., & Steiner, J. (2003). Measuring political deliberation: a discourse quality index. *Comparative European Politics*, 1(1), 21-48.

Stromer-Galley, J. (2007). Measuring deliberation's content: A coding scheme. *Journal of Public Deliberation*, 3(1), 12.

Trénel, M. (2004). Measuring the quality of online deliberation. Coding scheme 2.4. *Social Science Research Center Berlin, Germany*. Available at: http://www.wz-berlin.de/online-mediation/files/publications/quod_2_4.pdf .

Ullmann, T. D. (2004). maQ-Fragebogengenerator. Make a Questionnaire. Available online at: http://maq-online.de

Ullmann, T. D., Liddo, A. D., & Bachler, M. (2014). *Collective Intelligence Analytics Dashboard Usability Evaluation* (No. D4.6). The Open University. Retrieved from http://catalyst-fp7.eu/wp-content/uploads/2014/12/CATALYST_-D4.6.pdf

The CATALYST project has received funding from the European Union's Seventh Framework Programme for research, technological development and demonstration under grant agreement n°6611188

List of Figures

Figure 1 - Metric and alert server data flow 6
Figure 2 - UI for alerts in tree view 13
Figure 3 - Selection of goals for alerts 16
Figure 4 - Selection of exceptions for alerts 17
Figure 5 - Alert selection system 17
Figure 6 - Attention bias visualisation 18
Figure 7 - Rating bias visualisation 19
Figure 8 - Attention map visualisation 20
Figure 9 - Community interest visualisation 21
Figure 10 - Sub-communities network visualisation 22
Figure 11 - CI Dashboard alerts interface 23
Figure 12 - LiteMap alerts interface 24
Figure 13 - DebateHub alerts interface 25
Figure 14 - How often users ignore arguments when rating posts 32
Figure 15 - Simple deliberation map 33
Figure 16 - Balkanized rating histogram 35
Figure 17 - Eigenvector decomposition of uncorrelated interest space 35
Figure 18 - Eigenvector decomposition of balkanized interest space 36
Figure 19 - Topic tree color-coded by balkanized community 37

List of Tables

Table 1. List of metrics 7
Table 2. List of alerts 13
Table 3. Task performance 27
Table 4. Usability 28
Table 5. Uncorrelated ratings example 33
Table 6. Balkanized ratings example 34
Table 7. Eigenvector analysis 36
Table 8. Cluster ratings by harvesters 38
Table 9. Average silhouette score for cluster categories 38

The CATALYST project has received funding from the European Union's Seventh Framework Programme for research, technological development and demonstration under grant agreement n°6611188